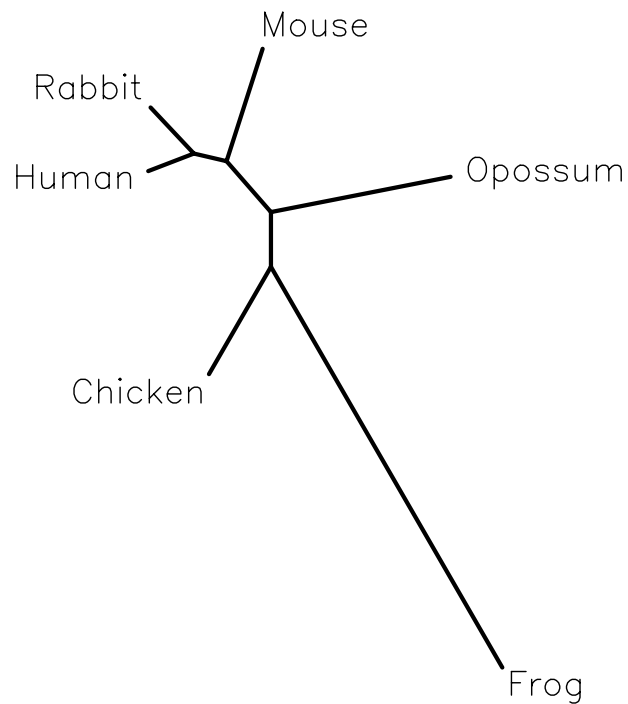

Detection of Recombination in Phylogenetic Data Sets with Hidden Markov Models

Dirk Husmeier and **Frank Wright**
Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK
Email: dirk@bioss.sari.ac.uk
Web: <http://www.bioss.sari.ac.uk/~dirk>

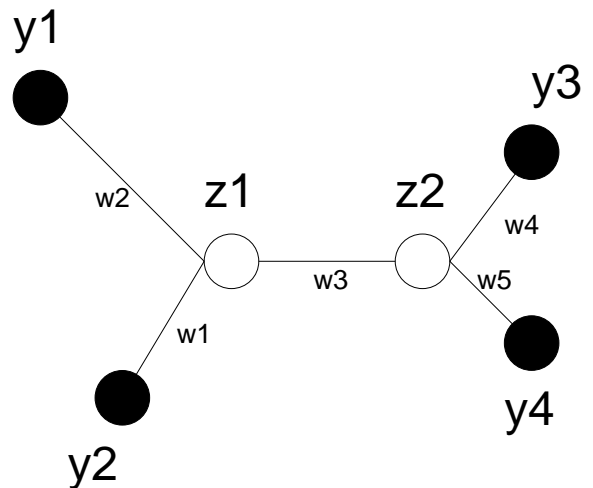
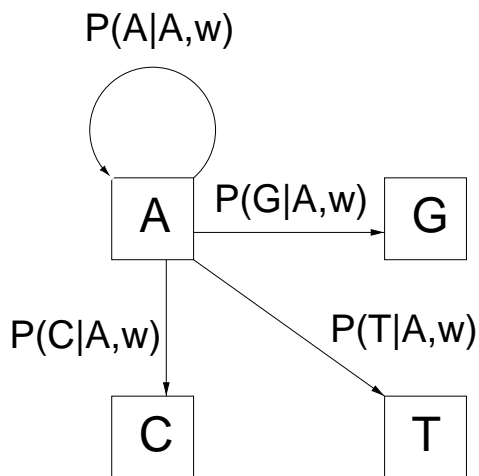
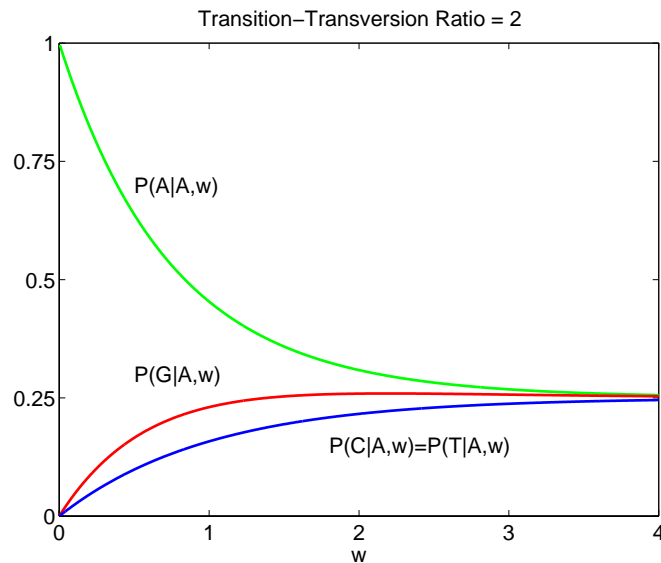
- Phylogeny
- Recombination
- Modelling Recombination with HMMs
- Simulation Studies

Phylogenetic Trees from DNA Sequence Alignments

Frog	GTCGCGGGTCAAACCTTTCCGTCTCGCG
Chicken	AGCATCGTTCTATTTTACCGGCTCCCG
Human	TGTATCGCTCAAGATTGCCATCGCGCG
Rabbit	TGTGTCGCTCAAGATTGCCATCGCGCG
Mouse	TGTCGTGGTCTAGATTGCCATCGCGCG
Opossum	TGTATCGCTCTAGTTTGCCAGCTCCCG



A Probabilistic Model of Evolution



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) = P(y_1 | z_1, w_2) P(y_2 | z_1, w_1) P(z_2 | z_1, w_3) P(y_3 | z_2, w_4) P(y_4 | z_2, w_5)$$

$$P(y_1, y_2, y_3, y_4 | \mathbf{w}) = \sum_{z_1} \sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

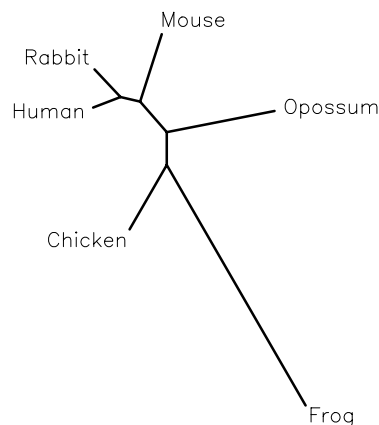
Probabilistic Approach to Phylogeny

Frog	GT C GCGGGTCAAAC TTTCCGTCTCGCG
Chicken	AG C ATCGTTCTATTTTACCGGCTCCCG
Human	TG T ATCGCTCAAGATTGCCATCGCGCG
Rabbit	TG T GTCGCTCAAGATTGCCATCGCGCG
Mouse	TG T CGTGGTCTAGATTGCCATCGCGCG
Opossum	TG T ATCGCTCTAGTTTGGCCAGCTCCCG

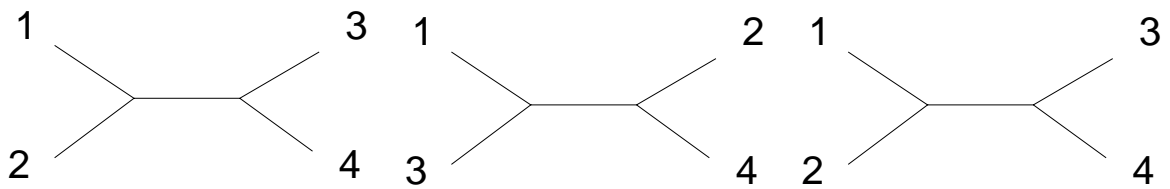
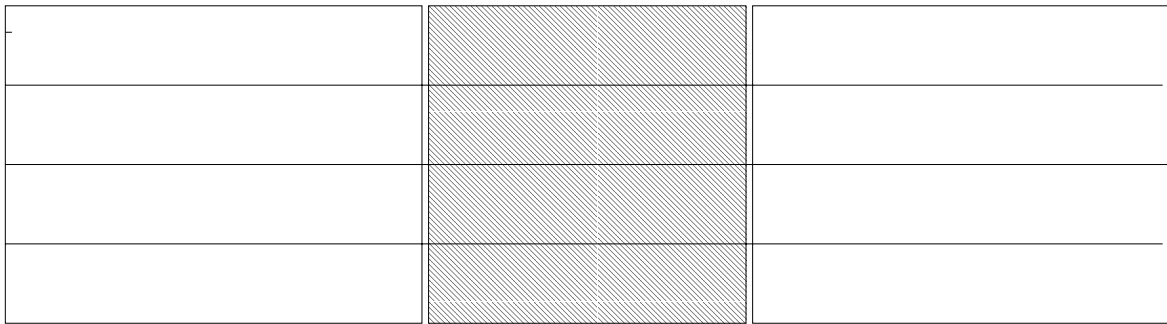
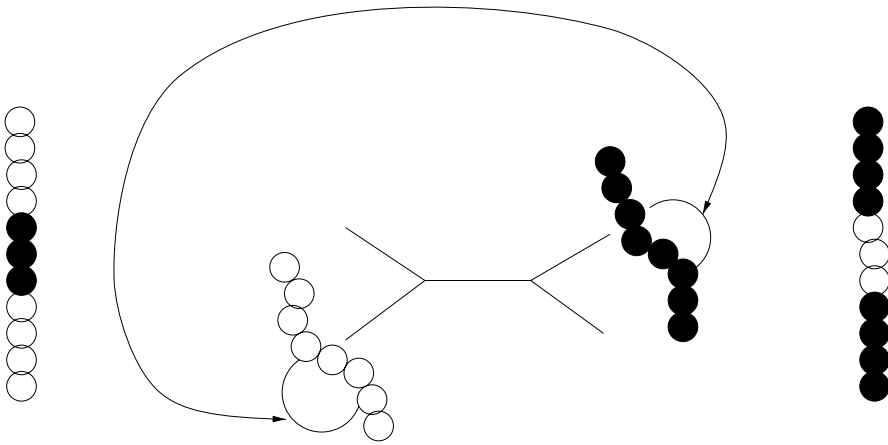
$$\mathbf{D} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$$

$$P(\mathbf{D}|\mathbf{w}, s) = \prod_{t=1}^N P(\mathbf{y}_t|\mathbf{w}, s)$$

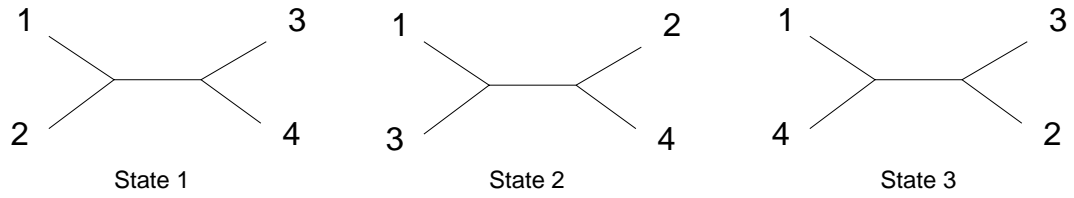
Optimisation of the **topology** s and the **branch lengths** \mathbf{w} by **maximum likelihood** .



Recombination



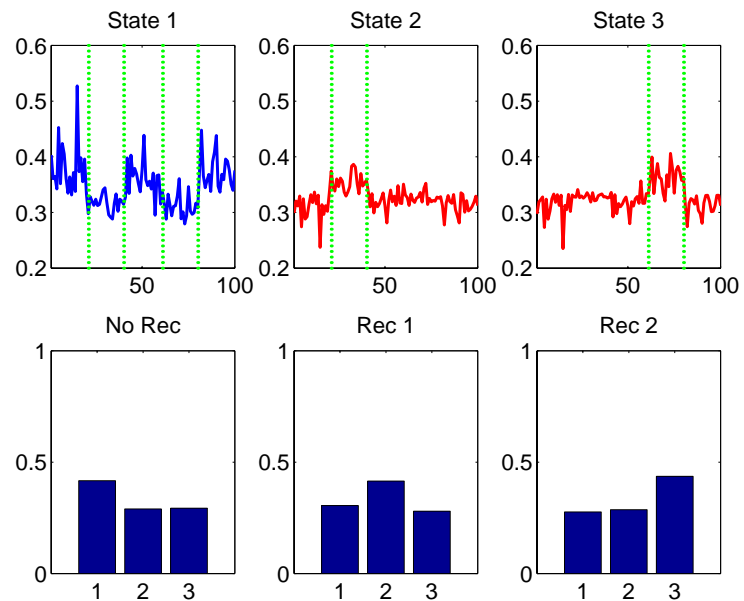
Naive approach



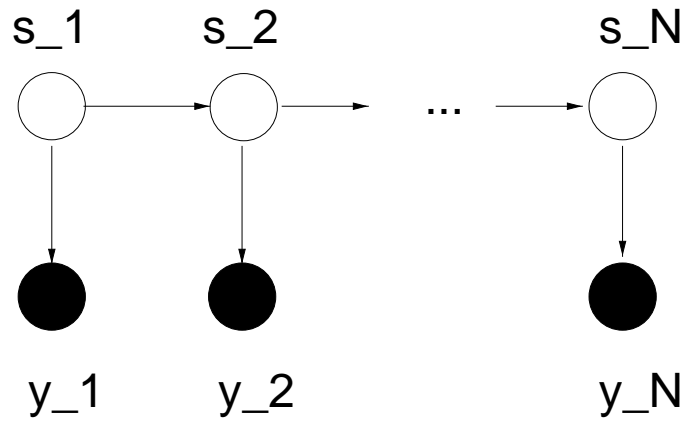
$$P(s_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | s_t) P(s_t)}{P(\mathbf{y}_t)}$$

$$P(s_t) = \text{Const}$$

$$P(s_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | s_t)}{\sum_{s'_t} P(\mathbf{y}_t | s'_t)}$$

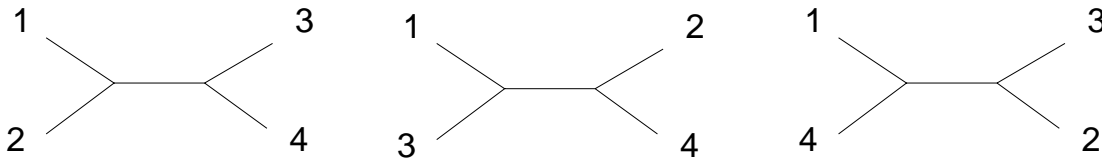


Modelling Recombination with HMMs I



$$P(\mathbf{y}_1, \dots, \mathbf{y}_N, s_1, \dots, s_N) = P(s_1) \prod_{t=2}^N P(s_t | s_{t-1}) \prod_{t=1}^N P(\mathbf{y}_t | s_t)$$

States s_t :

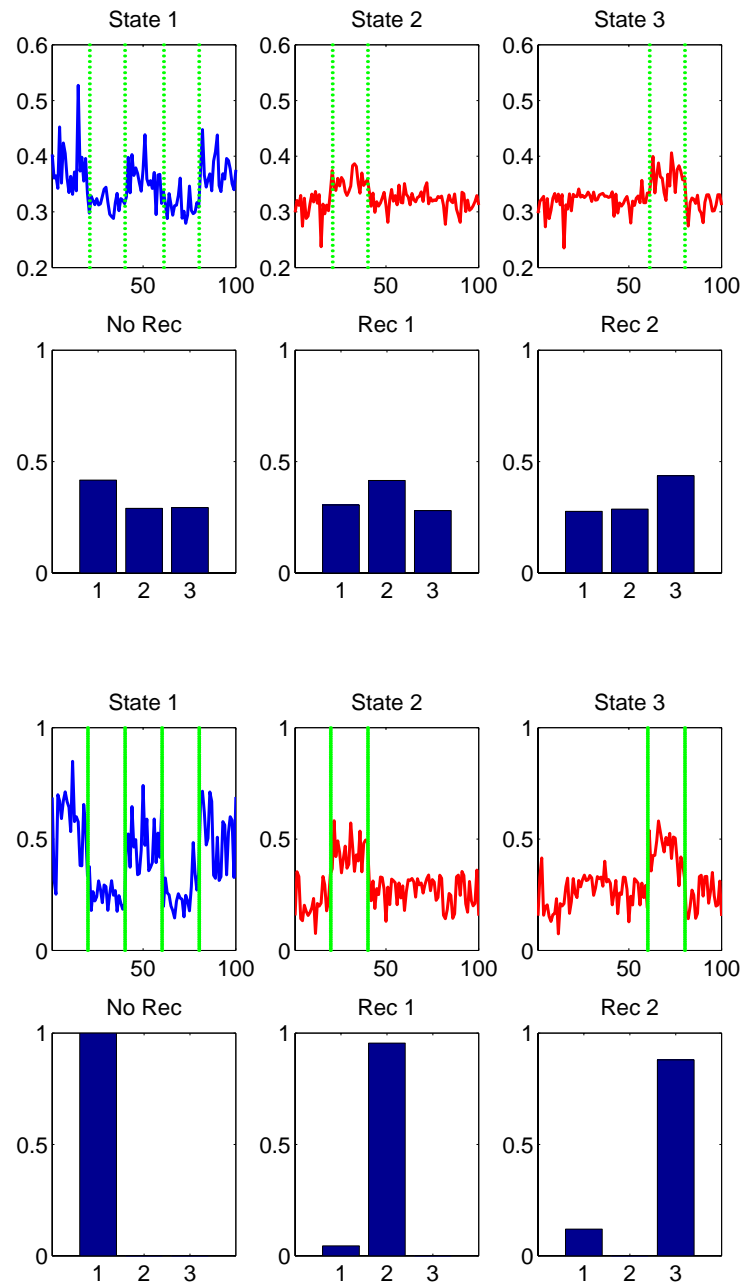


Transition Probabilities: $P(s_t | s_{t-1}) = \nu \delta_{s_t, s_{t-1}} + (1 - \delta_{s_t, s_{t-1}}) \frac{1 - \nu}{K - 1}$

State Sequence: $\mathbf{s} = (s_1, s_2, \dots, s_N)$

Mode of $P(\mathbf{s} | \mathbf{D}) \longrightarrow$ **Viterbi Path**

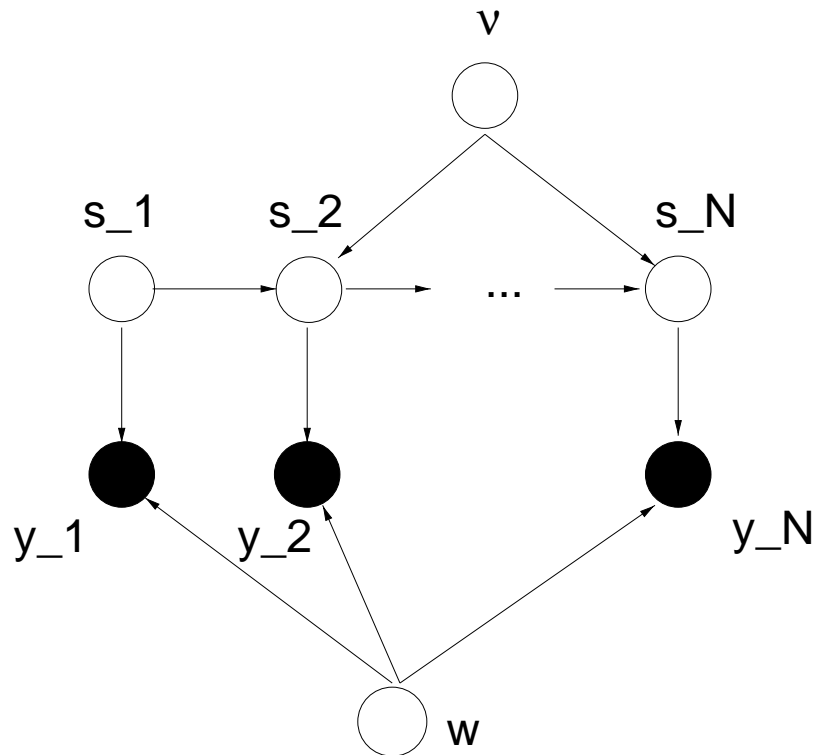
Comparison: Naive Approach versus HMM: $P(s_t|y_t)$



Modelling Recombination with HMMs II

Transition Probabilities: $P(s_t | s_{t-1}, \nu)$

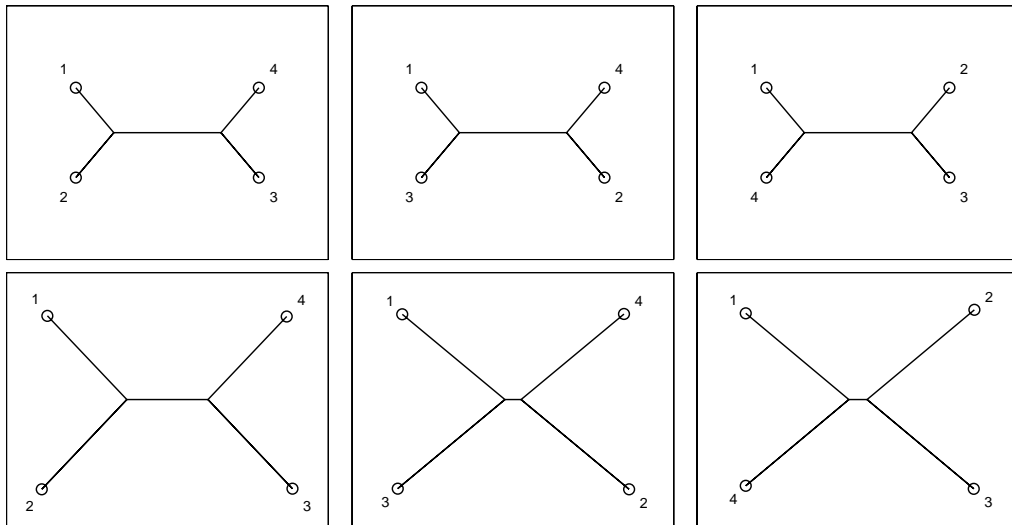
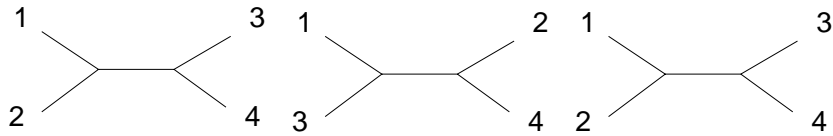
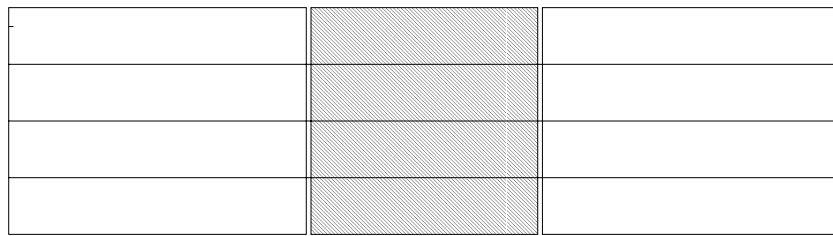
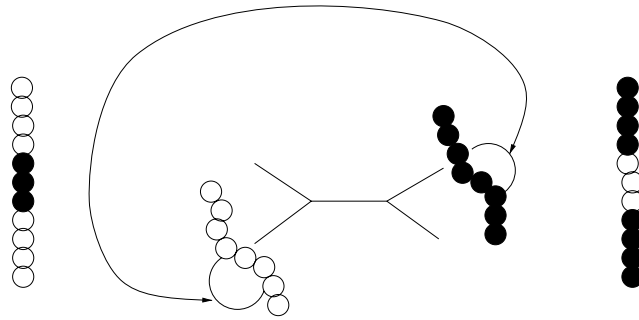
Emission Probabilities: $P(y_t | s_t, \mathbf{w})$



Previous work(McGuire): ν fixed, heuristic adaptation of \mathbf{w} .

Maximum Likelihood: $\operatorname{argmax} P(\mathbf{D} | \mathbf{w}, \nu)$

Determination of the Branch Lengths



EM Algorithm

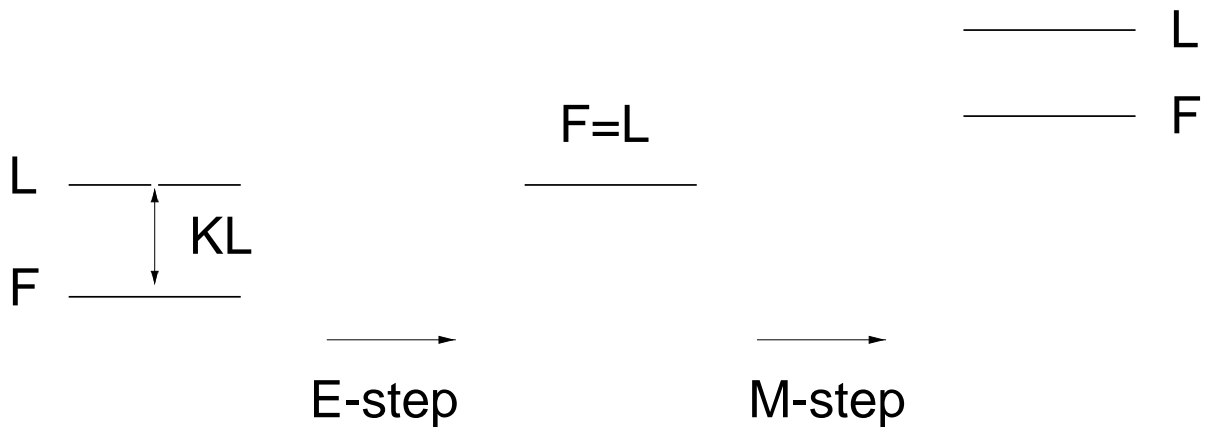
$$\begin{aligned} L(\mathbf{w}, \nu) &= \ln P(\mathbf{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{s}} P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu) \\ &= \ln \sum_{\mathbf{s}} \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})} Q(\mathbf{s}) \geq \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})} \end{aligned}$$

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})} = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)}{Q(\mathbf{s})} + \ln P(\mathbf{D}|\mathbf{w}, \nu)$$

$$L(\mathbf{w}, \nu) = F(\mathbf{w}, \nu) + KL[Q, P]$$

E-step $\longrightarrow Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)$

M-step \longrightarrow Maximise $F(\mathbf{w}, \nu)$



Application to Recombination

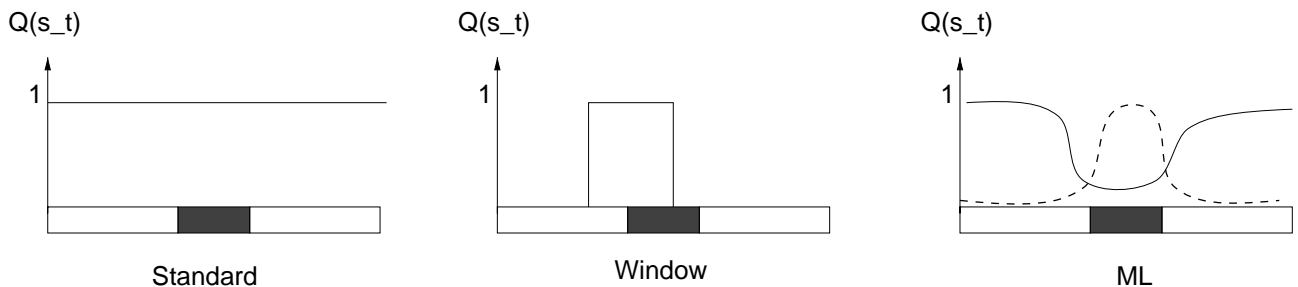
E-step: $Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)$

→ Forward-backward algorithm for HMMs

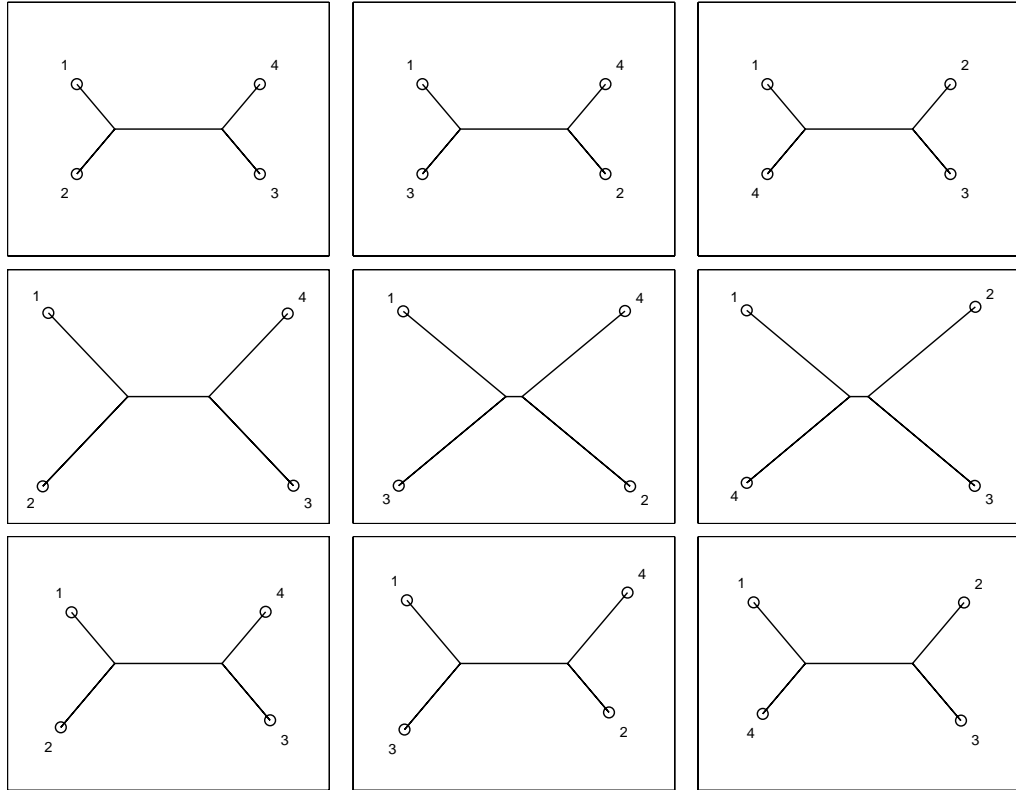
M-step: Maximise $F(\mathbf{w}, \nu) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)$

Recall:

$$P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu) = P(s_1) \prod_{t=2}^N P(s_t|s_{t-1}, \nu) \prod_{t=1}^N P(y_t|s_t, \mathbf{w})$$
$$\Rightarrow F(\mathbf{w}, \nu) = \sum_{t=1}^N \sum_{s_t} Q(s_t) \ln P(y_t|s_t, \mathbf{w})$$

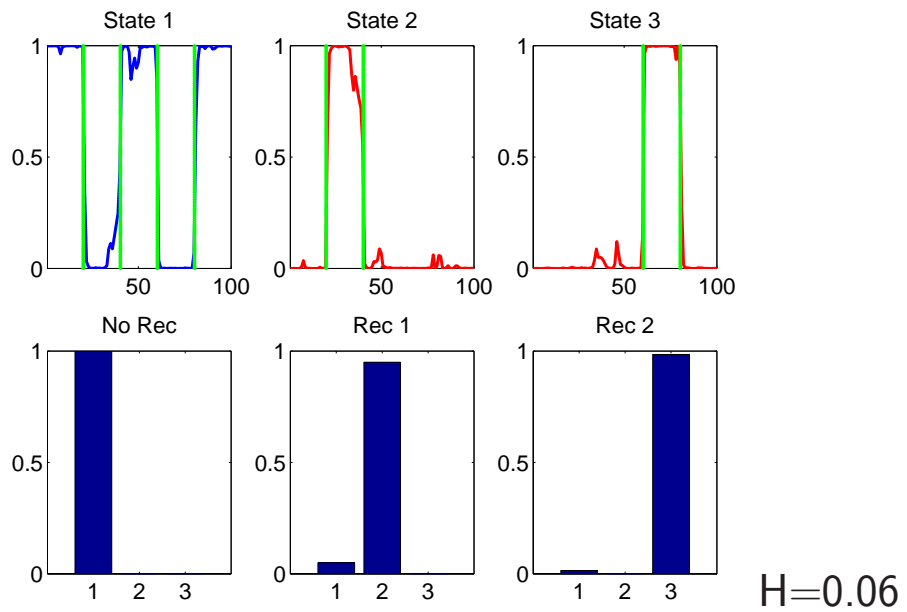
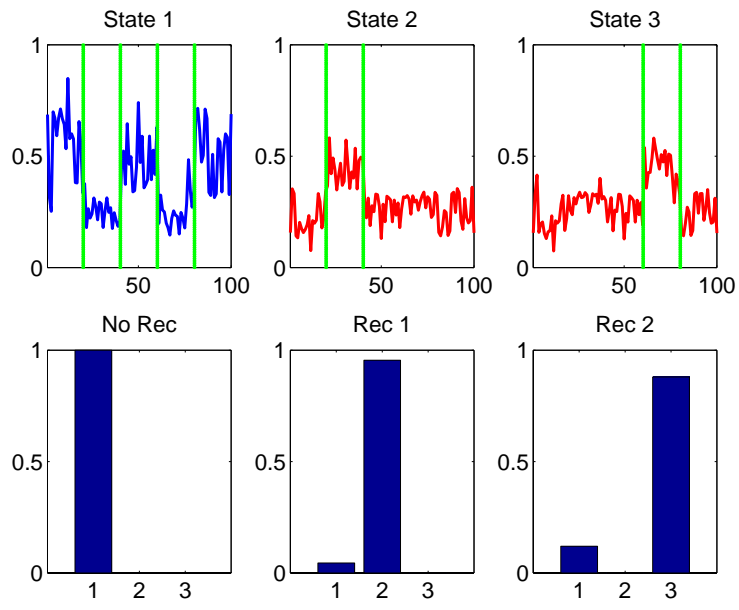


True and Predicted Trees



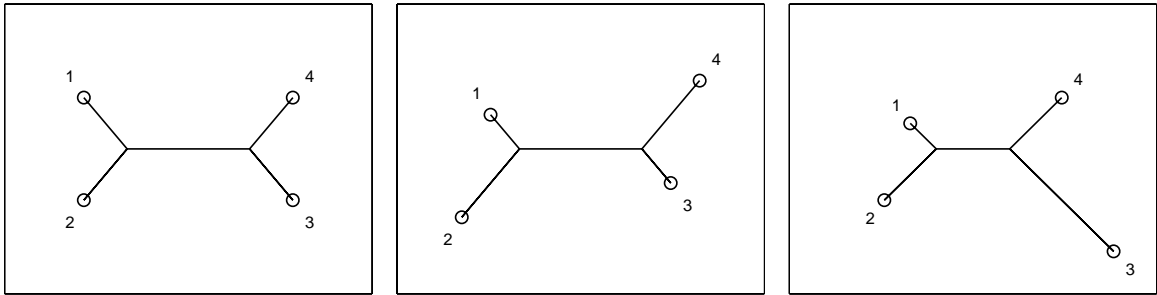
Training method	State 1	State 2	State 3
HMM CML	0.13	0.23	0.22
HMM EM	0.04	0.06	0.04

Comparison of HMMs: CML versus ML: $P(s_t|y_t)$



Simulation Experiment

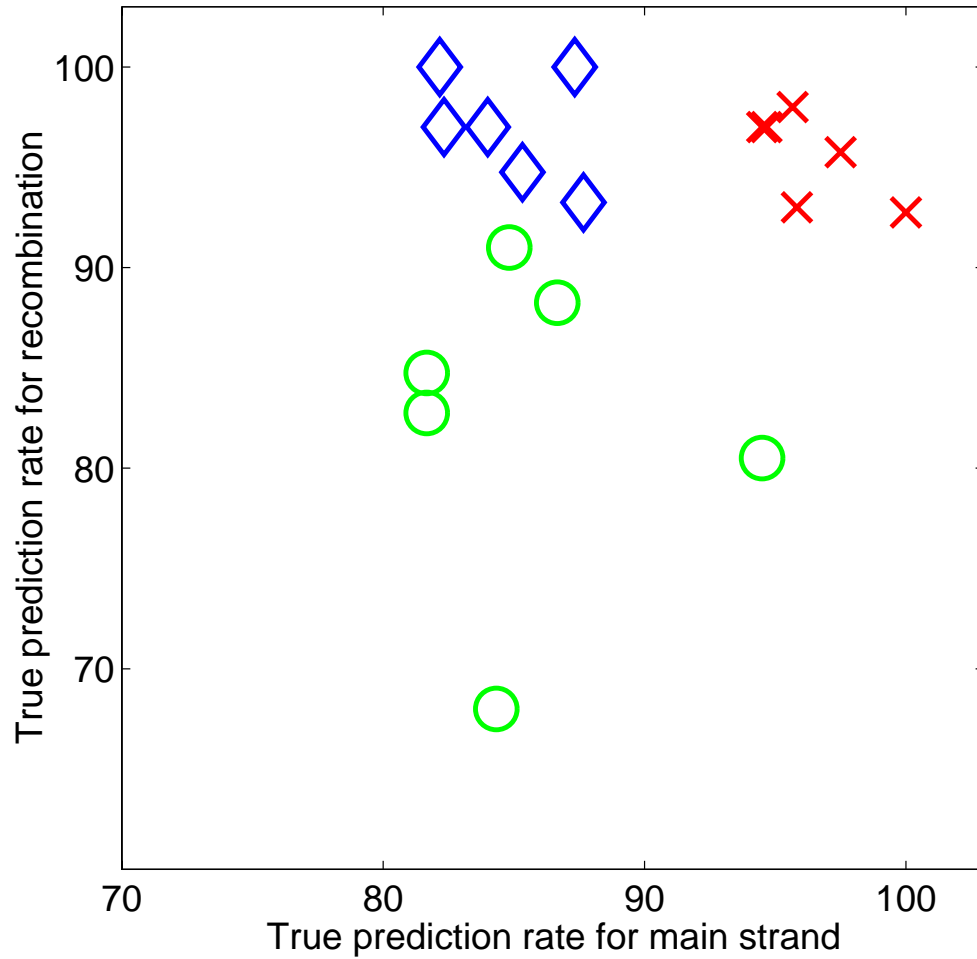
True phylogenetic trees



Recombination events

	Recombination Event	Affected Sites
Exp 1	$s_2 \rightarrow s_3$	201-400
	$s_2 \rightarrow s_4$	601-800
Exp 2	$s_2 \rightarrow s_3$	201-300
	$s_2 \rightarrow s_4$	601-900

Results 1: Classification Performance



Crosses EM
Diamonds CML, $\nu = 0.8$
Circles CML, $\nu = 0.6$

Results 2: Classification Entropy

$$H = -\frac{1}{N \ln K} \sum_{t=1}^N \sum_{s_t=1}^K P(s_t | \mathbf{y}_t) \ln P(s_t | \mathbf{y}_t)$$

$H = 1$ Random classifier

$H = 0$ Perfect classifier

	EM	CML-06	CML-08
$\mathbf{w}_1, 200, 200$	0.04	0.87	0.75
$\mathbf{w}_2, 200, 200$	0.06	0.86	0.72
$\mathbf{w}_3, 200, 200$	0.10	0.93	0.86
$\mathbf{w}_1, 100, 300$	0.06	0.86	0.74
$\mathbf{w}_2, 100, 300$	0.06	0.85	0.73
$\mathbf{w}_3, 100, 300$	0.08	0.92	0.85

Summary

- Modelling recombination with HMMs
- Earlier approach (McGuire):
 - Constrained maximum likelihood
 - Window method
 - ν fixed
- New approach:
 - Maximum likelihood with the EM algorithm

<http://www.bioss.sari.ac.uk/~dirk>